



---

## IBM Data Stage Lab Guide#12

---

### Remove Duplicate on Datastage 8.5

---

#### **Description:**

*BISP is committed to provide BEST learning material to the beginners and advance learners. In the same series, we have prepared a list of beginner's guide and FAQs for IBM Data Stage. We have built complete financial Data Model and various data transformation techniques. Download many such learning documents, student guide, Lab Guide and Hands-on practice materials. Join our professional training.*

#### **History:**

Version	Description Change	Author	Publish Date
0.1	Initial Draft	Varun Khare	12th Aug 2012
0.1	Review#1	Amit Sharma	18 <sup>th</sup> Aug 2012

# Mapping Sheet

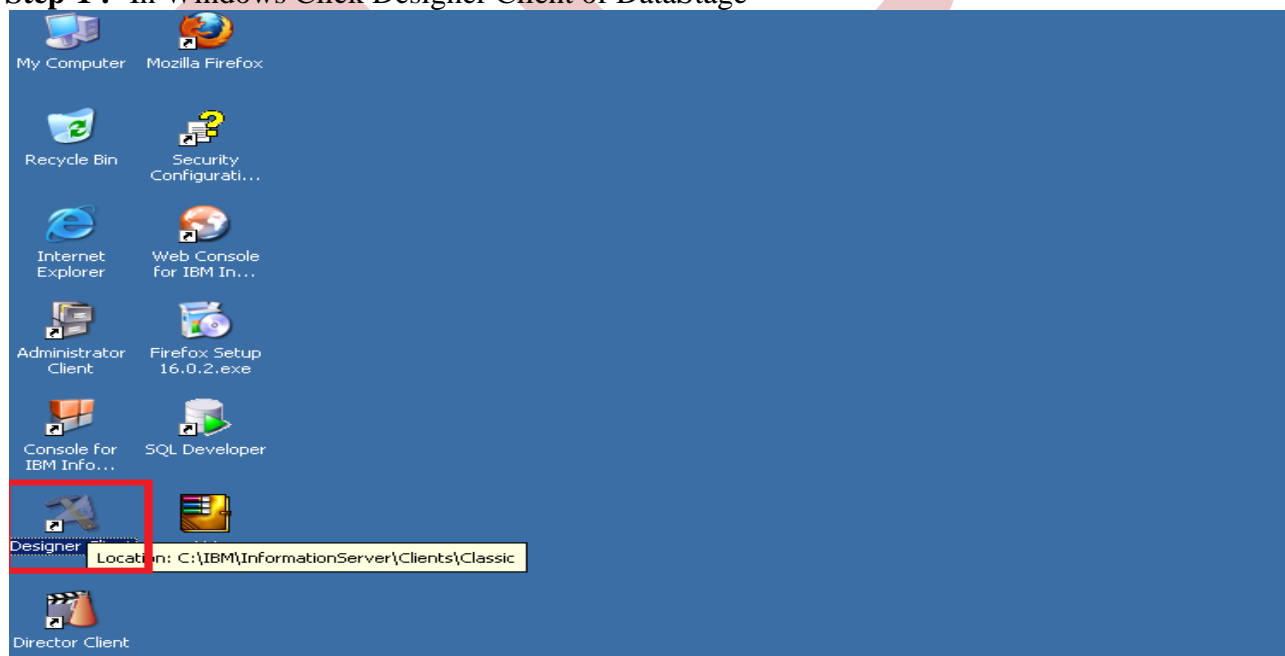
## Remove Duplicate

### Source- Flat File

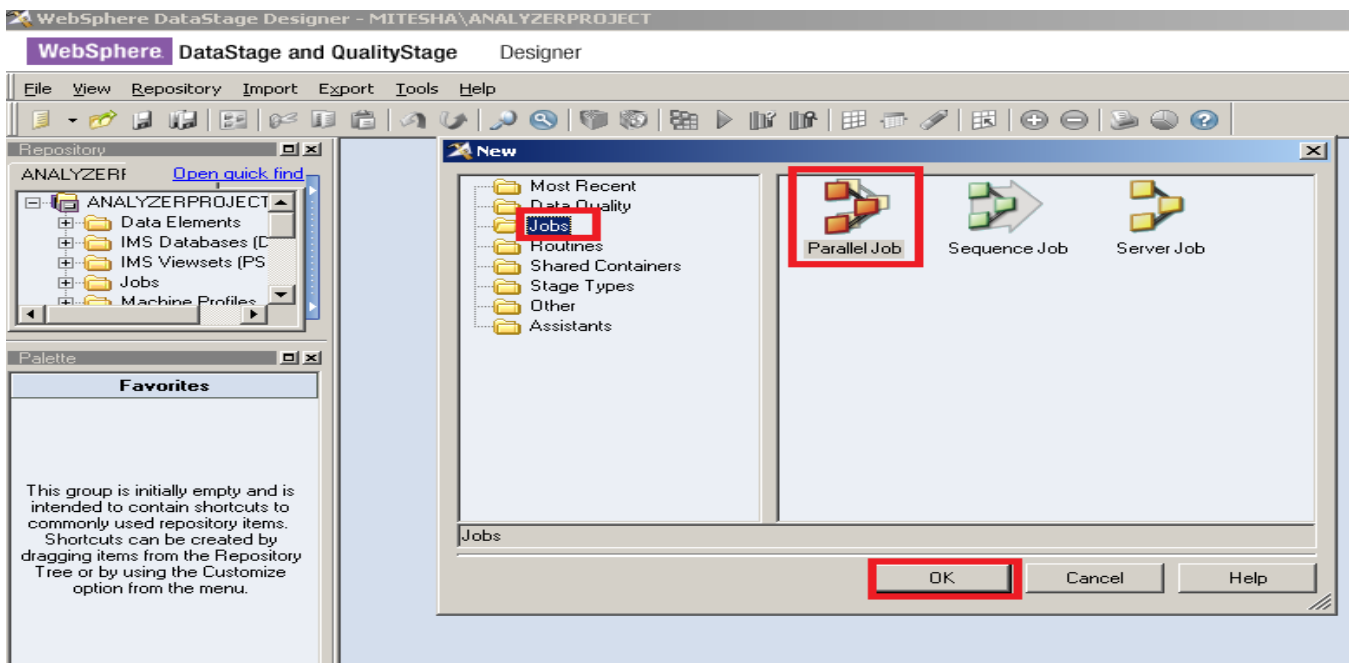
### Target- Oracle

Transformation	Remove Duplicate Stage						
Source Name	Flat File						
Target Name	Oracle						
Flat File Source Details				Oracle Details			
Entity Name	Data Types	Is Null	Expression	Destination Entity Name	DestinationField Name	Is Null	Data Types
ACCOUNT_OFFICER_CD	NUMBER(5)	N	select distinct s.* from SRC_ACCOUNT_OFFICER_CD s;	src_creditcard	ACCOUNT_OFFICER_CD	N	NUMBER(5)
CREATED_BY	VARCHAR2(30)	y		src_creditcard	CREATED_BY	y	VARCHAR2(30)
CREATION_DATE	VARCHAR2(100)	y		src_creditcard	CREATION_DATE	y	VARCHAR2(100)
DEFINITION_LANGUAGE	VARCHAR2(4)	y		src_creditcard	DEFINITION_LANGUAGE	y	VARCHAR2(4)
ENABLED_FLAG	VARCHAR2(1)	y		src_creditcard	ENABLED_FLAG	y	VARCHAR2(1)
LAST_MODIFIED_BY	VARCHAR2(30)	y		src_creditcard	LAST_MODIFIED_BY	y	VARCHAR2(30)
LAST_MODIFIED_DATE	VARCHAR2(100)	y		src_creditcard	LAST_MODIFIED_DATE	y	VARCHAR2(100)
LEAF_ONLY_FLAG	VARCHAR2(1)	y		src_creditcard	LEAF_ONLY_FLAG	y	VARCHAR2(1)
ACCOUNT_OFFICER_DISPLAY_CD	VARCHAR2(10)	y		src_creditcard	ACCOUNT_OFFICER_DISPLAY_CD	y	VARCHAR2(10)

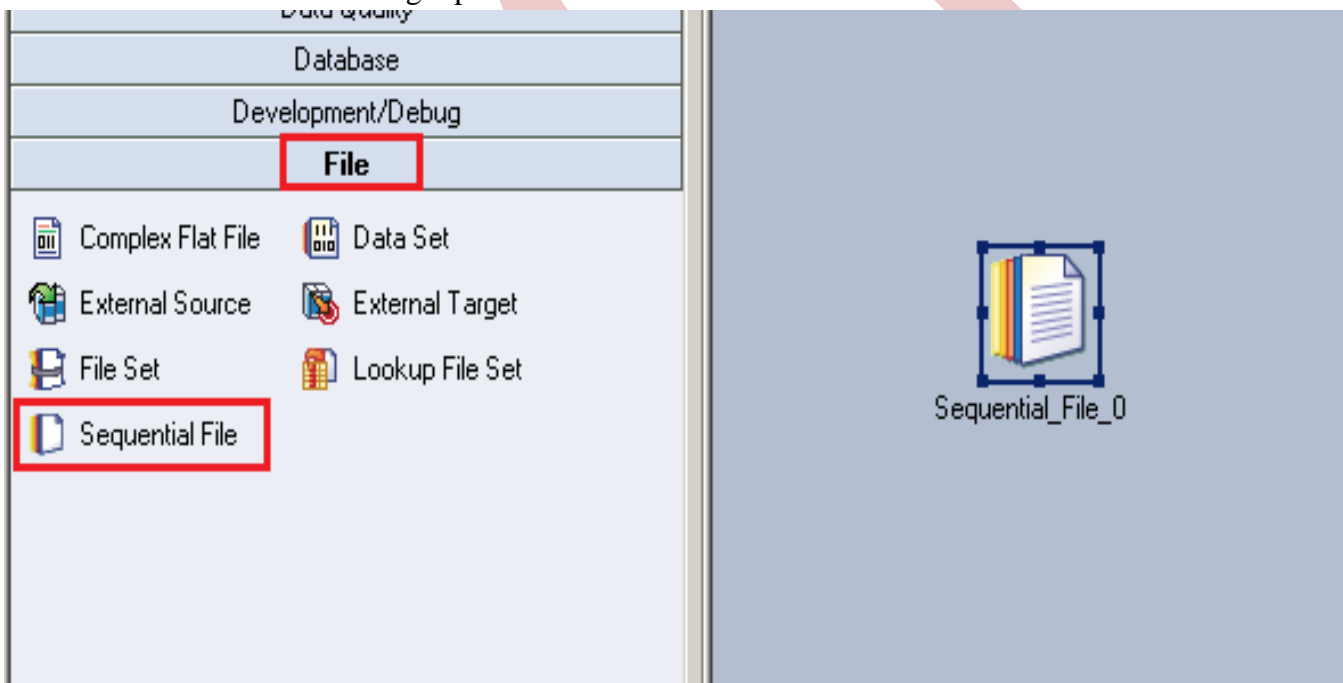
**Step-1 :-** In Windows Click Designer Client of DataStage



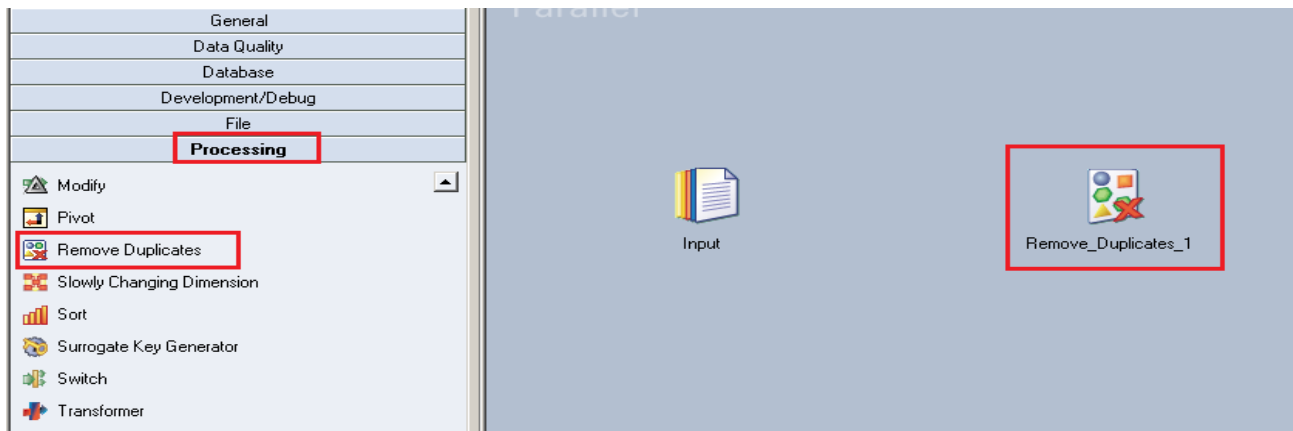
**Step-2 :-** New Window opens, then click Jobs ,then click Parallel Job because we are using parallel jobs.



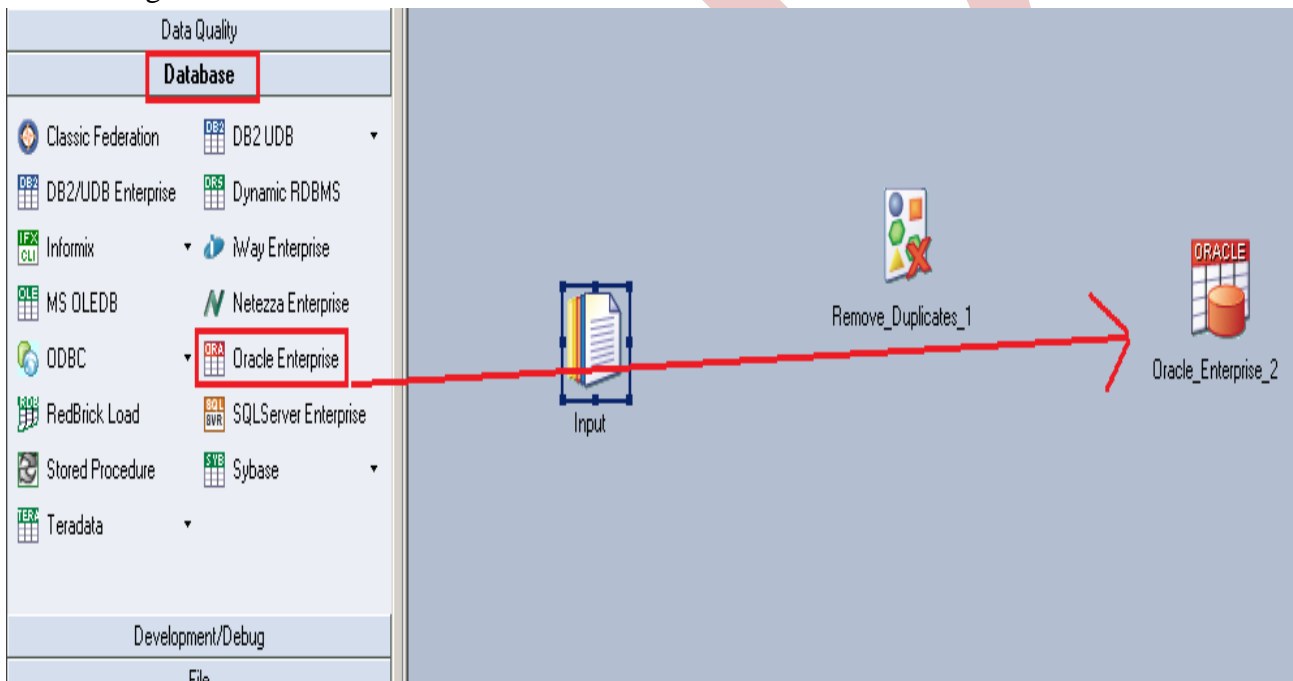
**Step-3 :-** First go on File Palette then scroll the cursor and choose Sequential File drag it to Parallel pane and rename it. This is used for taking input from flat file.



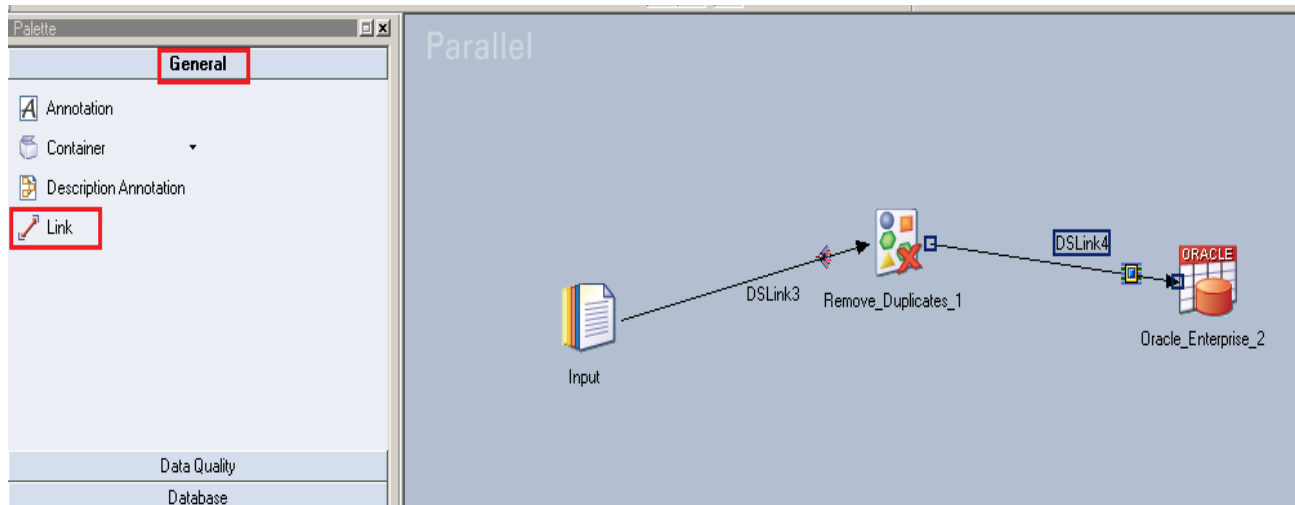
**Step-4 :-** Now go to processing palette and choose remove duplicate stage as we know that this stage is useful for load unique value only and remove duplicates value from source file.



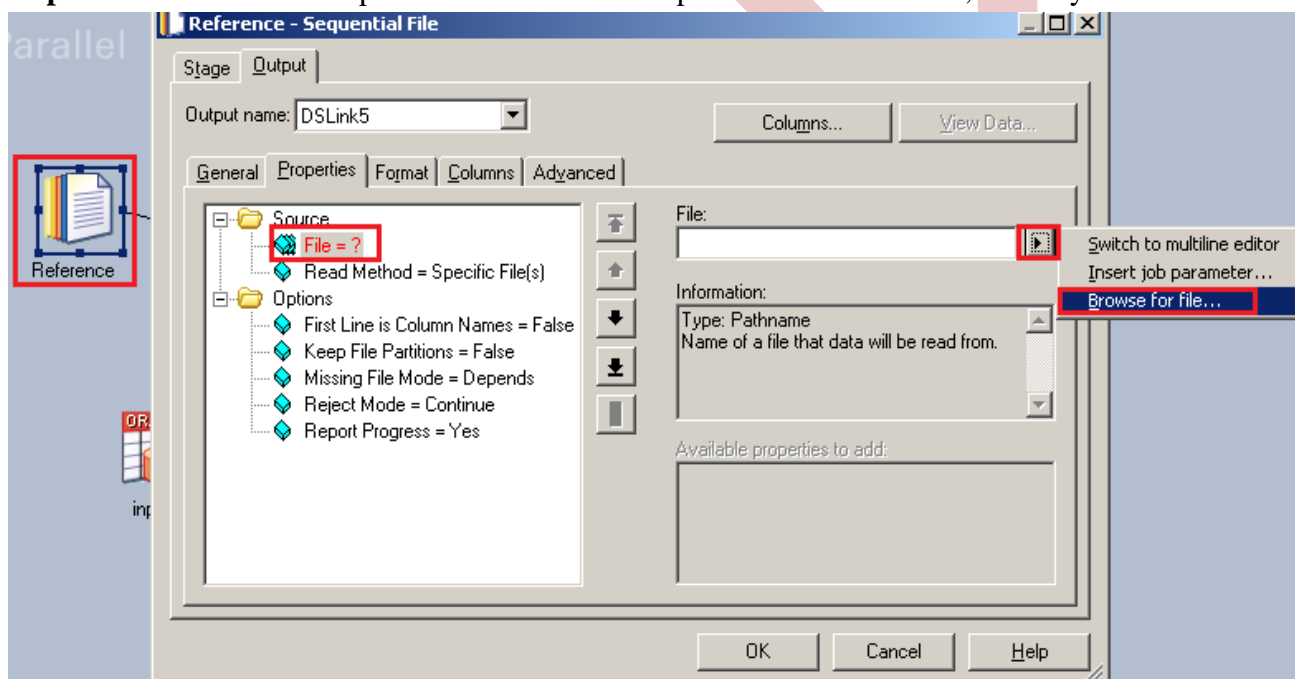
**Step-5 :-** Choose Oracle Enterprise from database. Database Palette and drag it to parallel pane. This is used for save target data into oracle.



**Step-6:-** Now make an connection with these stages that's why we have to use this link stage for connecting these stages.

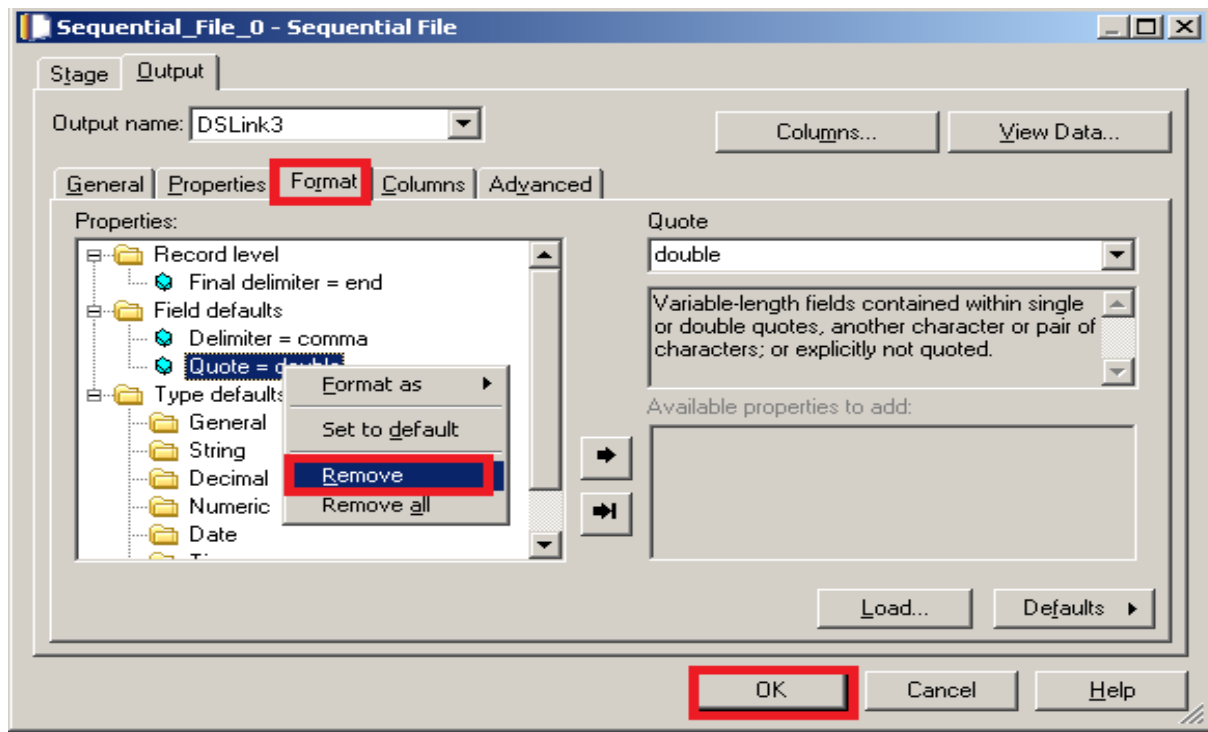


**Step-7:-** Double click on Input file and browse an input file that can be .txt, .csv any test file.

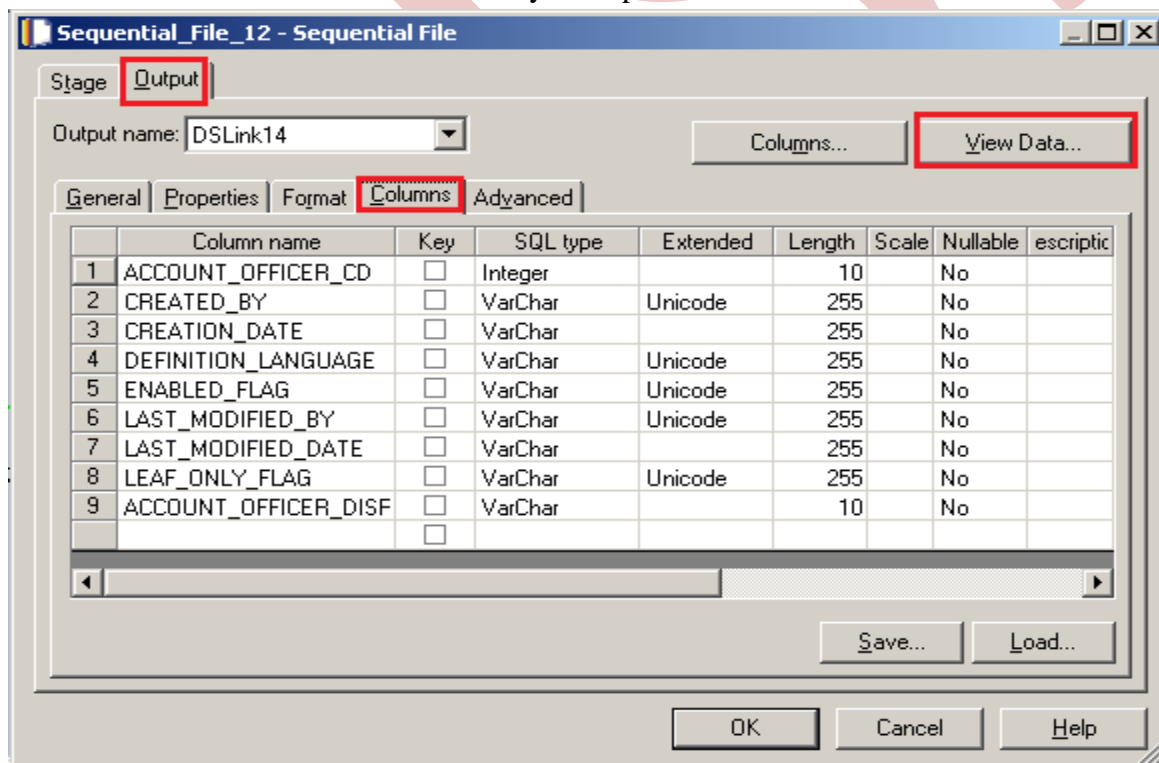


**Step-8:-** Go to Format tab and Remove double Quote because we don't need them then Click on OK.

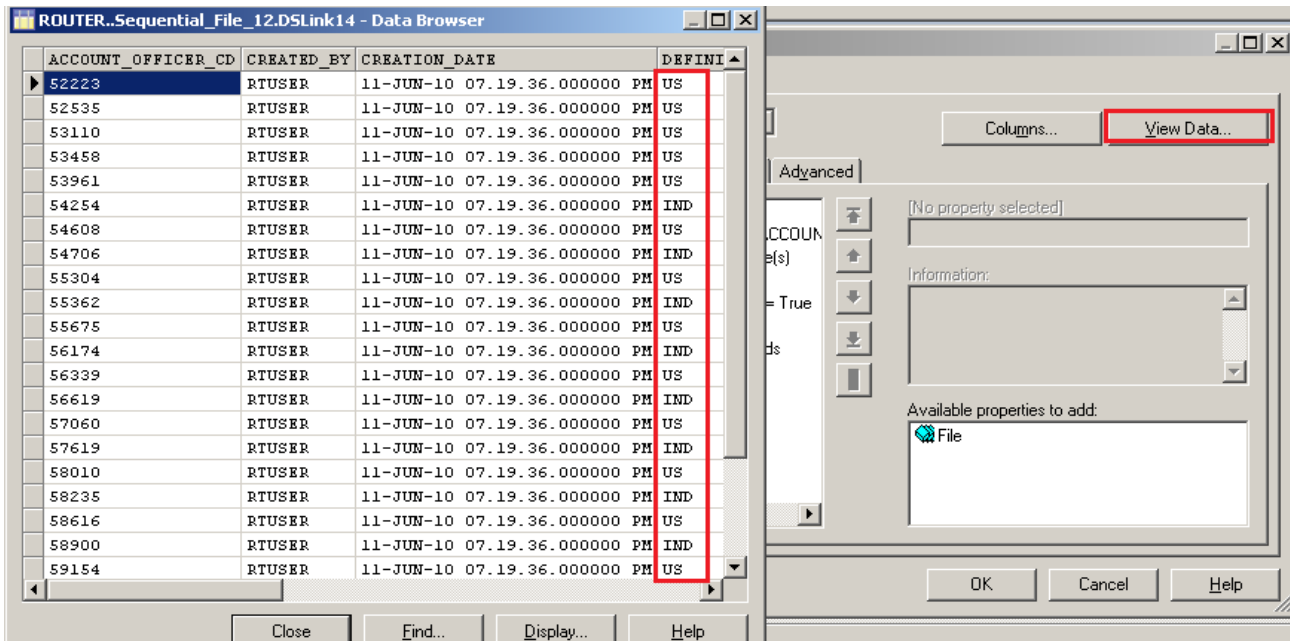
Note:- If you use date as a input than specify here which type of format you used



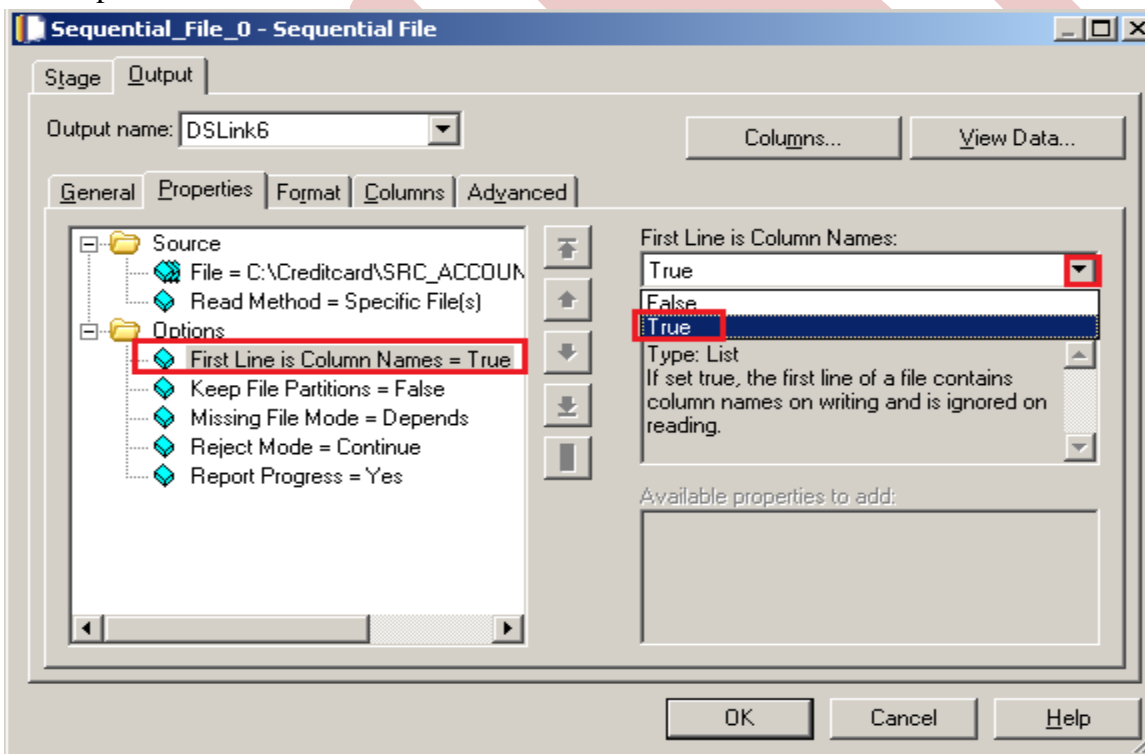
**Step-9:-** Go to Columns tab and enter same column names as declared in Input file and put data type, length than click on View Data. This will show your Input Data.



**Step-10:-** Now click on View Data and Ok. If your data is shown like this that means you have made successful connection between input file and Datastage otherwise it will give some error like column mismatch error.



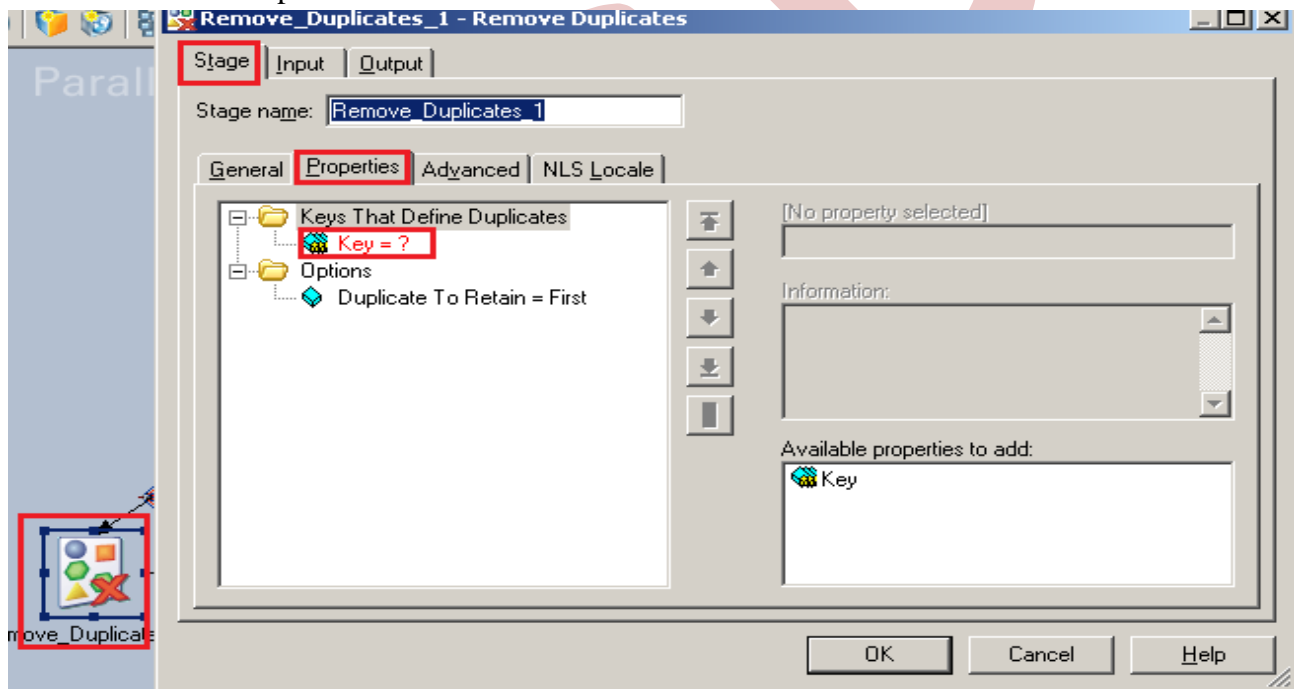
**Step-11:-** After choosing file, select 'True' from First line in column Names dropdown for removing first line from Input file.



**Step-12 :-** Make input file and remember column name

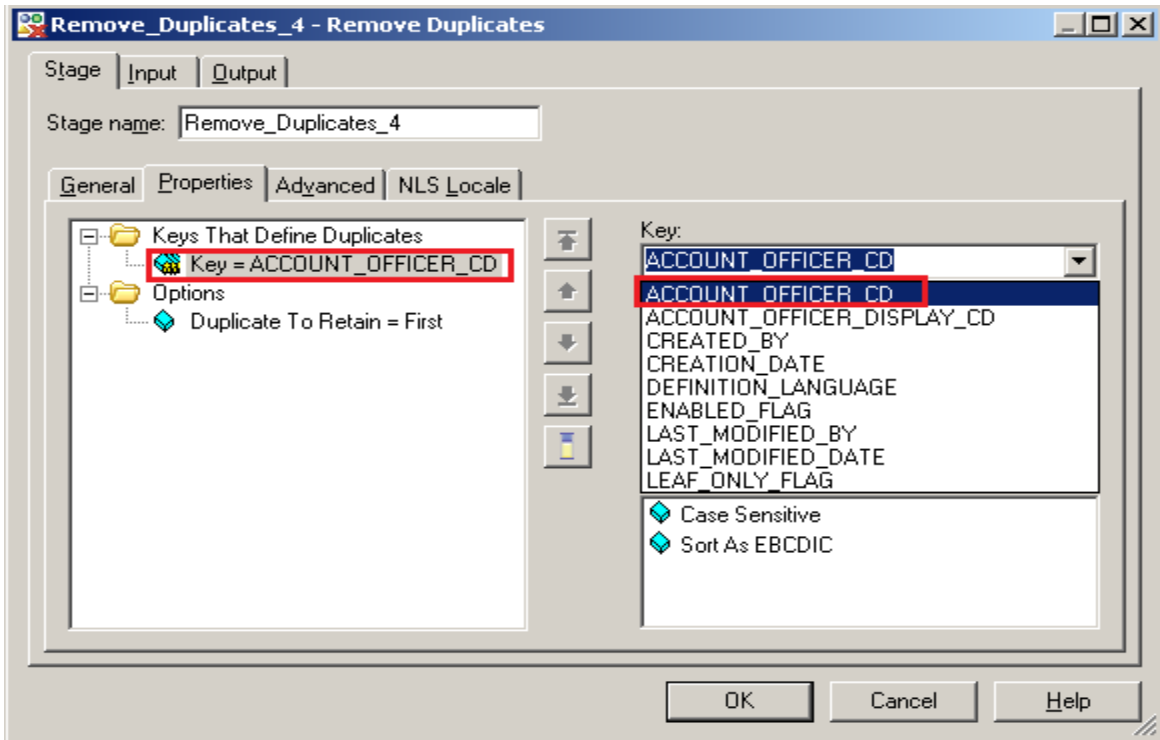
ACCOUNT_OFFICER_CD	CREATED_BY	CREATION_DATE	DEFINITION_LANGUAGE	ENABLED_FLAG	LAST_MODIFIED_BY	LAST_MODIFIED_DATE	LEAF_ONLY_FLAG	ACCOUNT_OFFICER_DISPLAY_CD
52223	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	52223.36,
52535	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	52534.89,
53110	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	53110.24,
53458	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	53457.6,
53961	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	53960.7,
54254	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	54253.88,
54608	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	54607.99,
54706	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	54705.91,
55304	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	55303.81,
55362	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	55362.01,
55675	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	55674.97,
56174	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	56173.79,
56339	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	56338.61,
56619	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	56618.92,
57060	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	57059.61,
57619	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	57619.17,
58010	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	58010.18,
58235	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	58235.35,
58616	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	58616.48,
58900	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	58899.61,
59154	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	59153.7,
59567	RTUSER	11-JUN-10	07.19.36.000000	PM,IND,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	59566.99,
60254	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	60254.29,
60740	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	60740.15,
60951	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	60951.47,
61134	RTUSER	11-JUN-10	07.19.36.000000	PM,US,Y	RTUSER,11-JUN-10	07.19.36.000000	PM,Y	61133.62,

**Step-13:-** Now double click on remove duplicate stage and see these properties here simple properties are available the main thing is there's key used that means we have to choose unique column from source file where we want that operation.

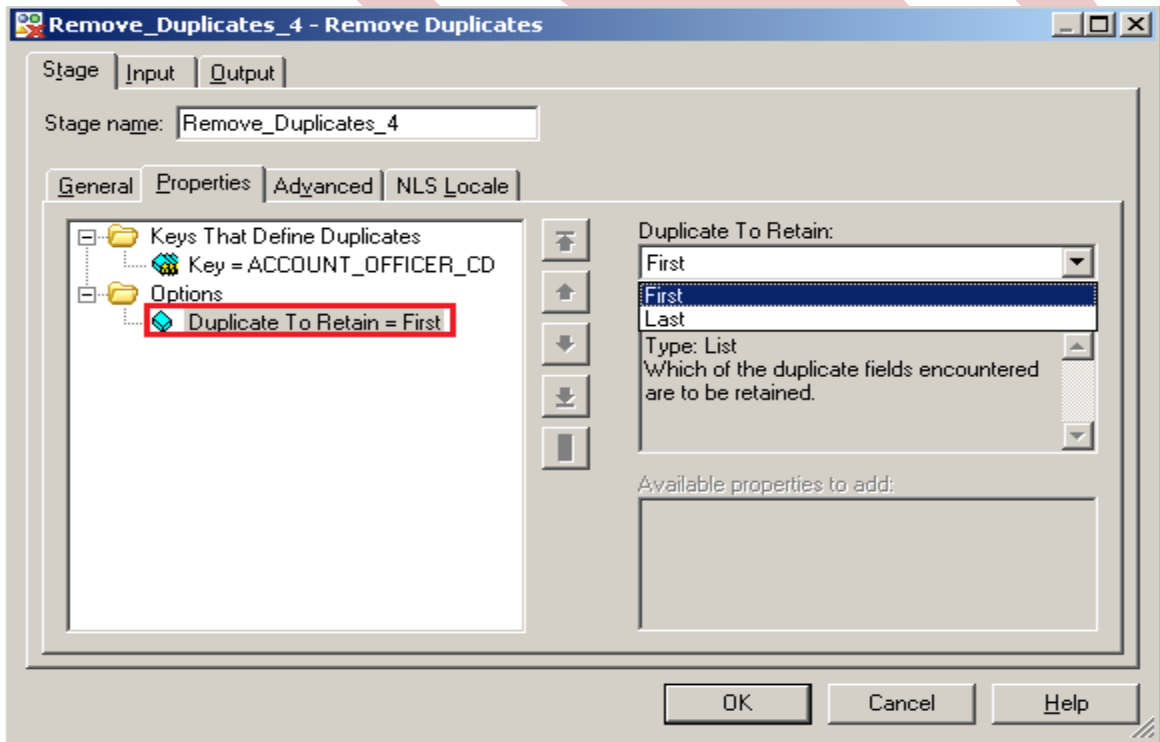


**Step-14:-** See here we choose this column we want to remove duplicate data based on this column values if in this column when repeated values available than remove that row.

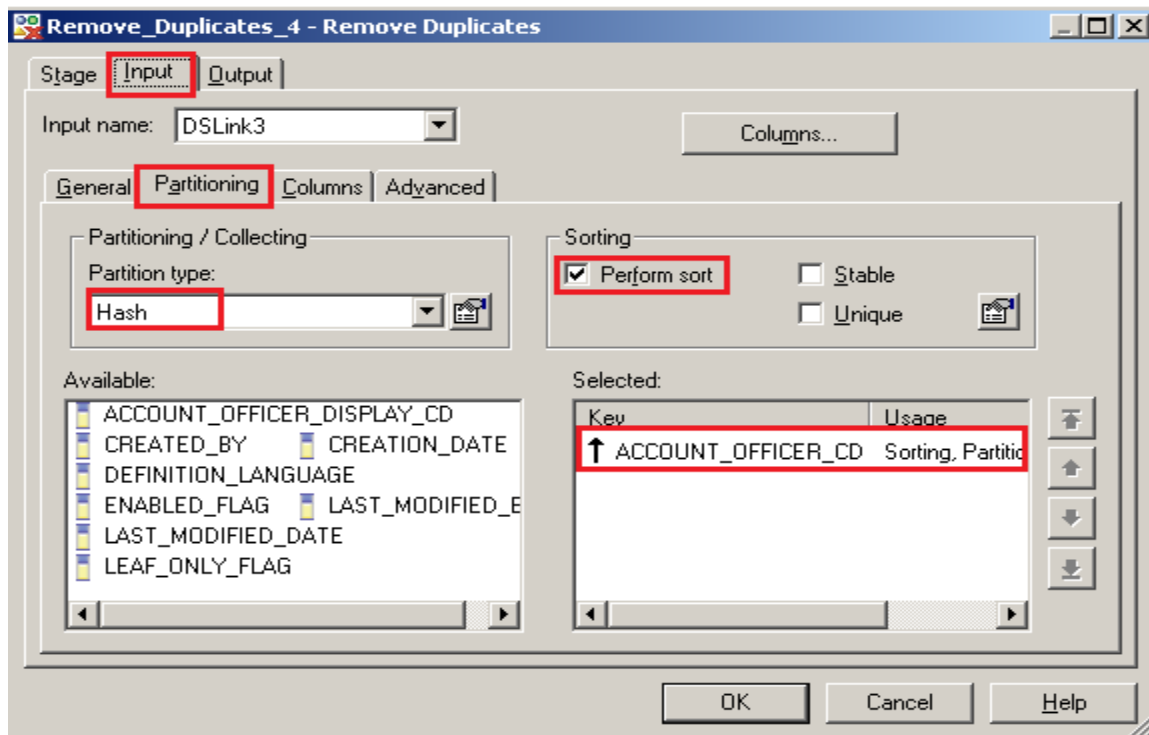




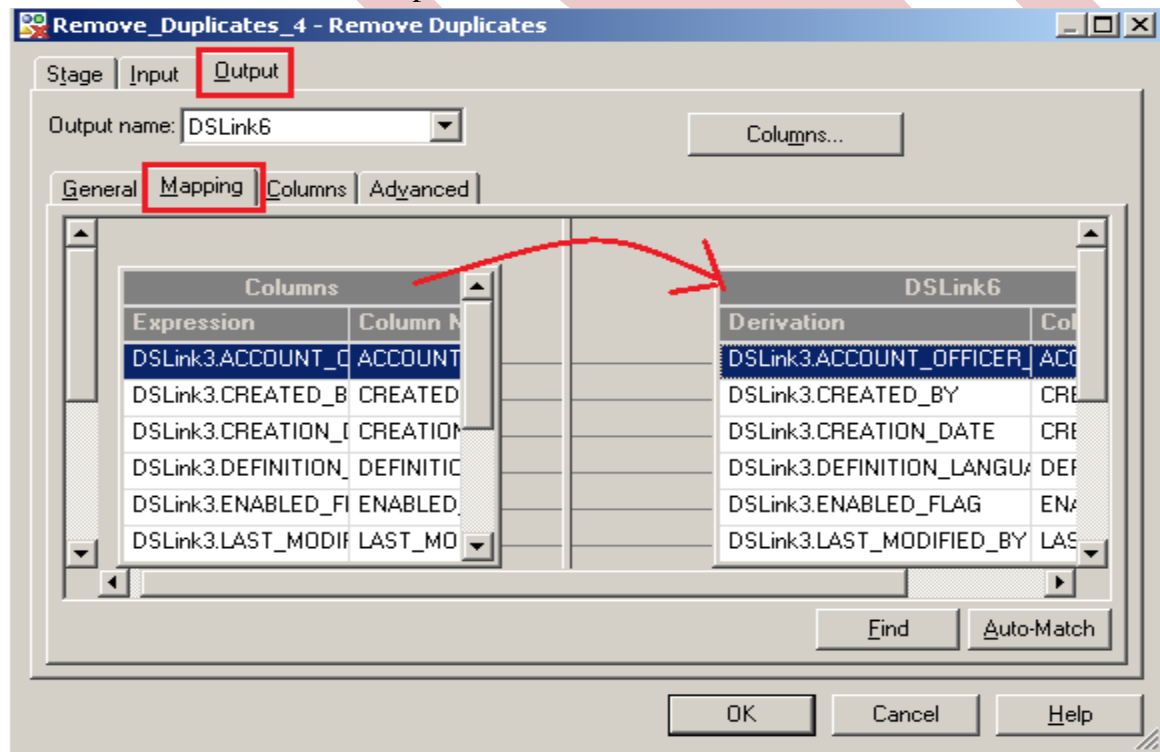
**Step-15:-** In this option which row we want to remove from source data this specify here that means here we choose if two rows were same than first row eliminate or last one we have to specify this things here.



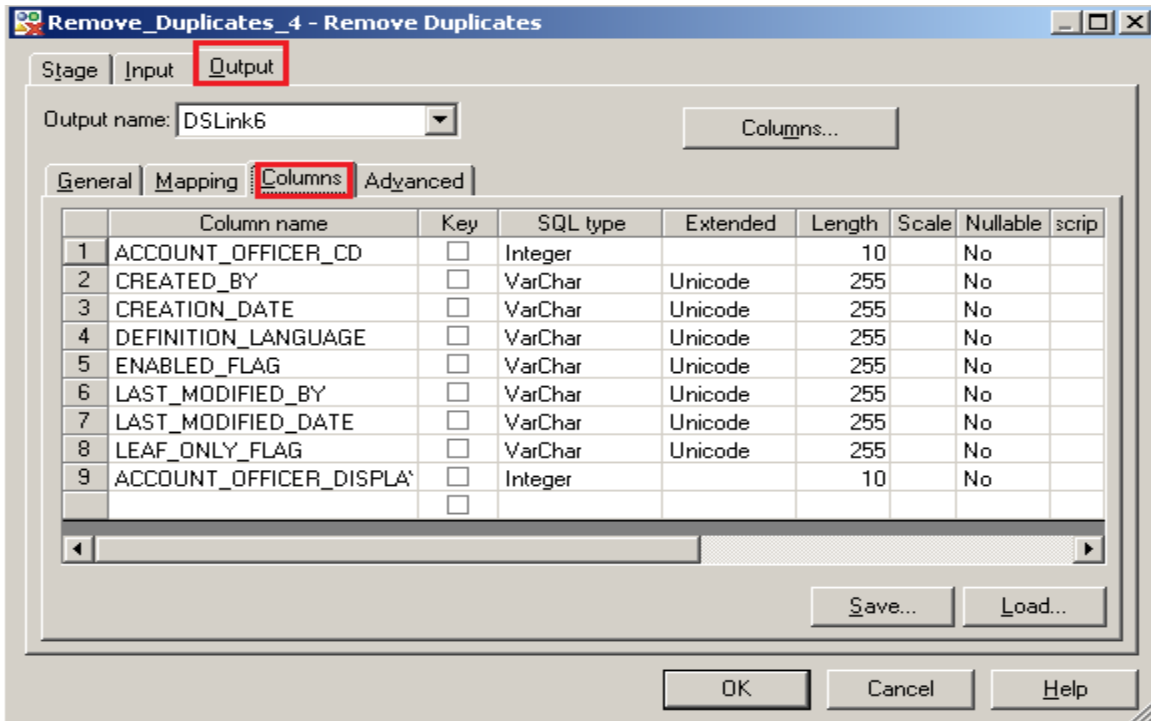
**Step-16:-** Now click to input and choose hash key partition because these are key based stage so we have to choose hash and make an unique key.



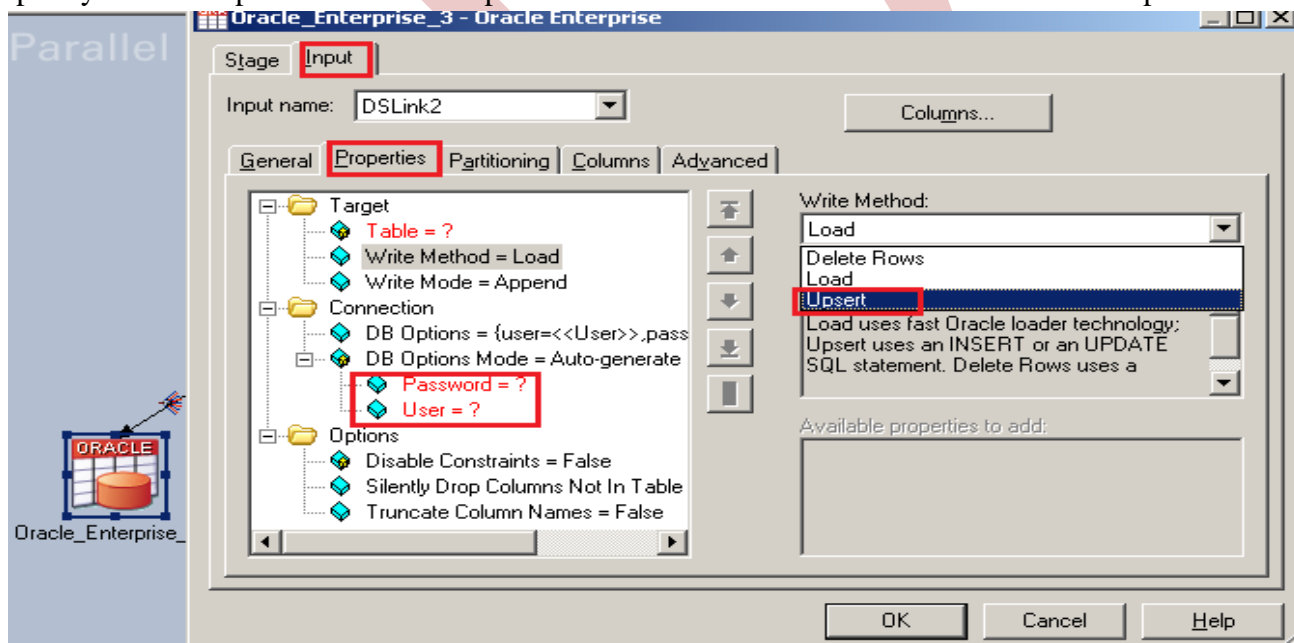
**Step-17:-** Now go to output tab than select all links and drag them into Output link for sending data which satisfies filter condition into output links..



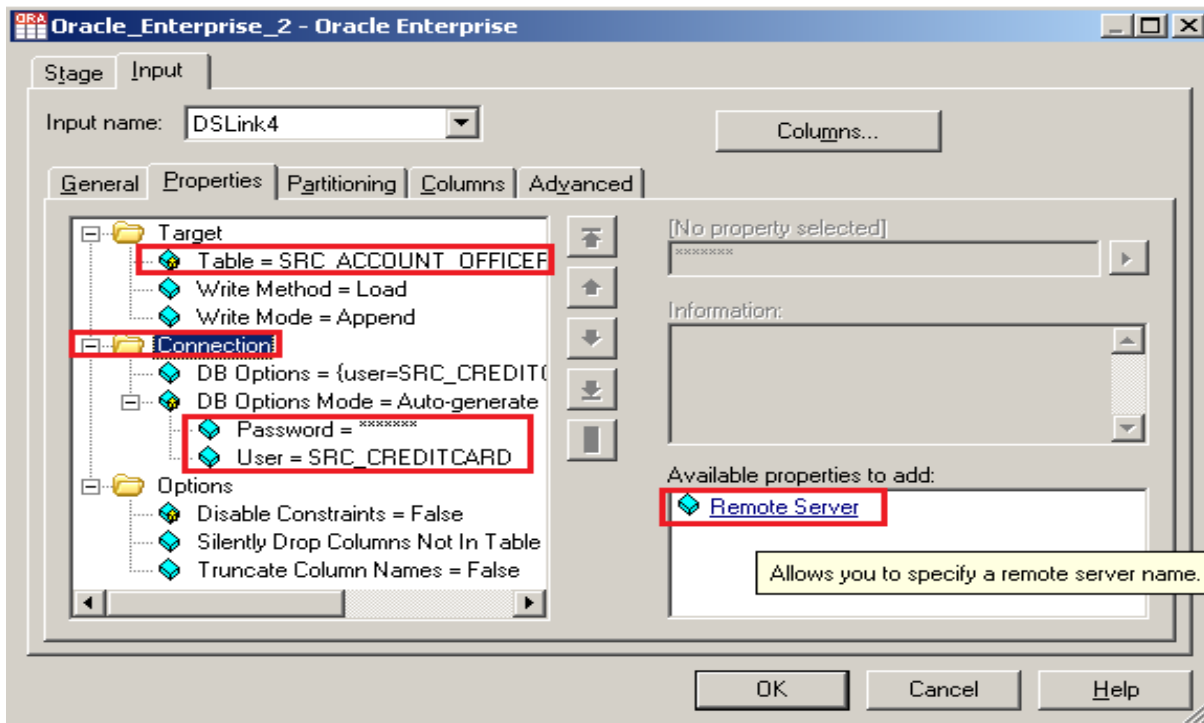
**Step-18:-** See here in output column there is one more column available and this column is useful for see the values of changing operation.



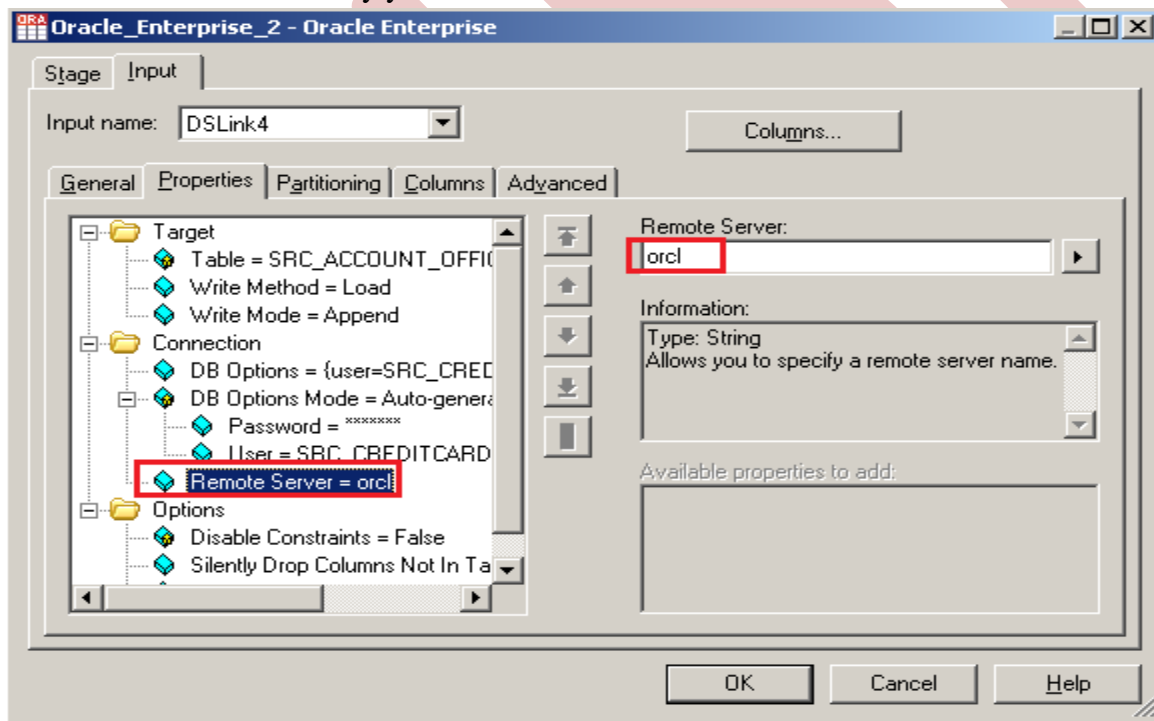
**Step-19 :-** Now double click on Oracle enterprise stage than it looks like it and we have to enter our table name here in which table we want to insert our data and there multiple options if we directly want to load data than we simple used write method as load otherwise we manually put queries on it and also we have to specify username/password on it this password should be matched with oracle username/password.



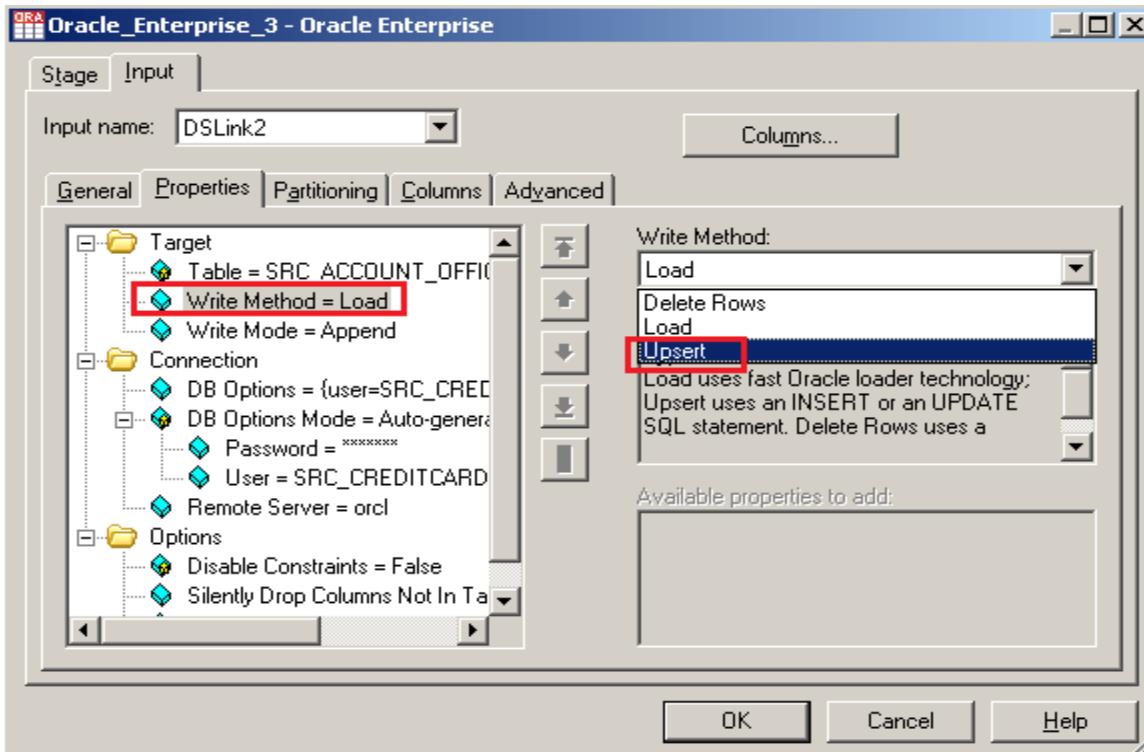
**Step-20:-** There's one more option here once you click on connection tab then remote server options are popup then we have to specify our server name of oracle.



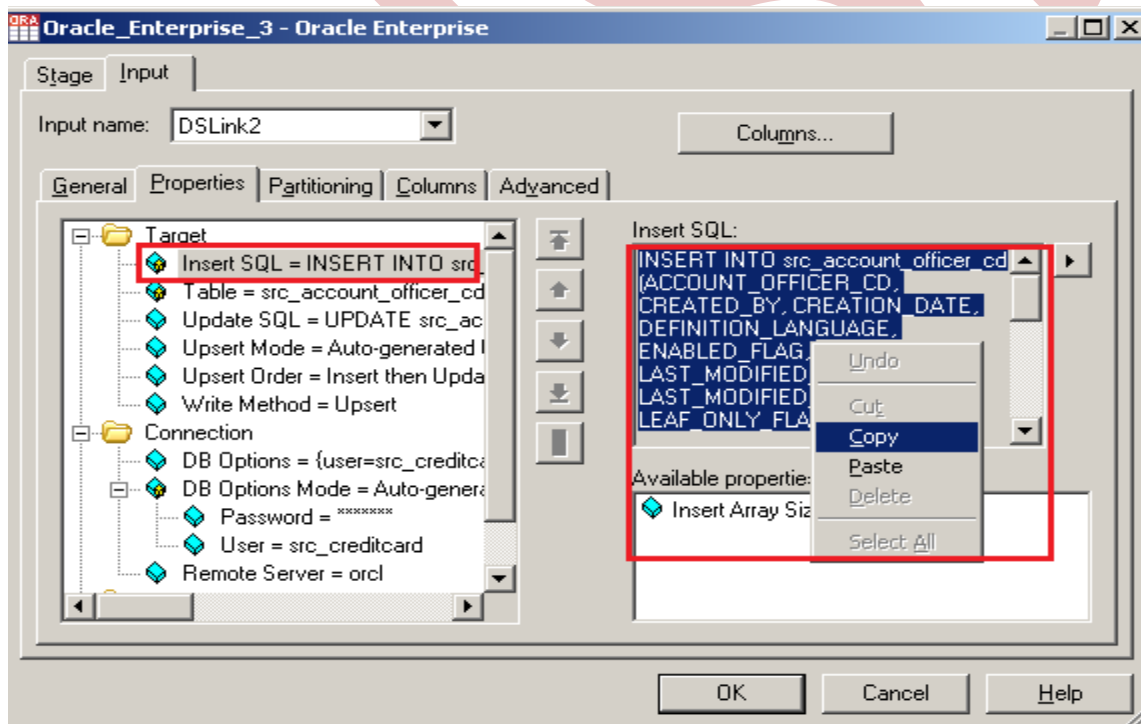
**Step-21:-** See below, my server name is ORCL and username/password and table name. Ensure that you put correct information then only you will be able to load the data into table.



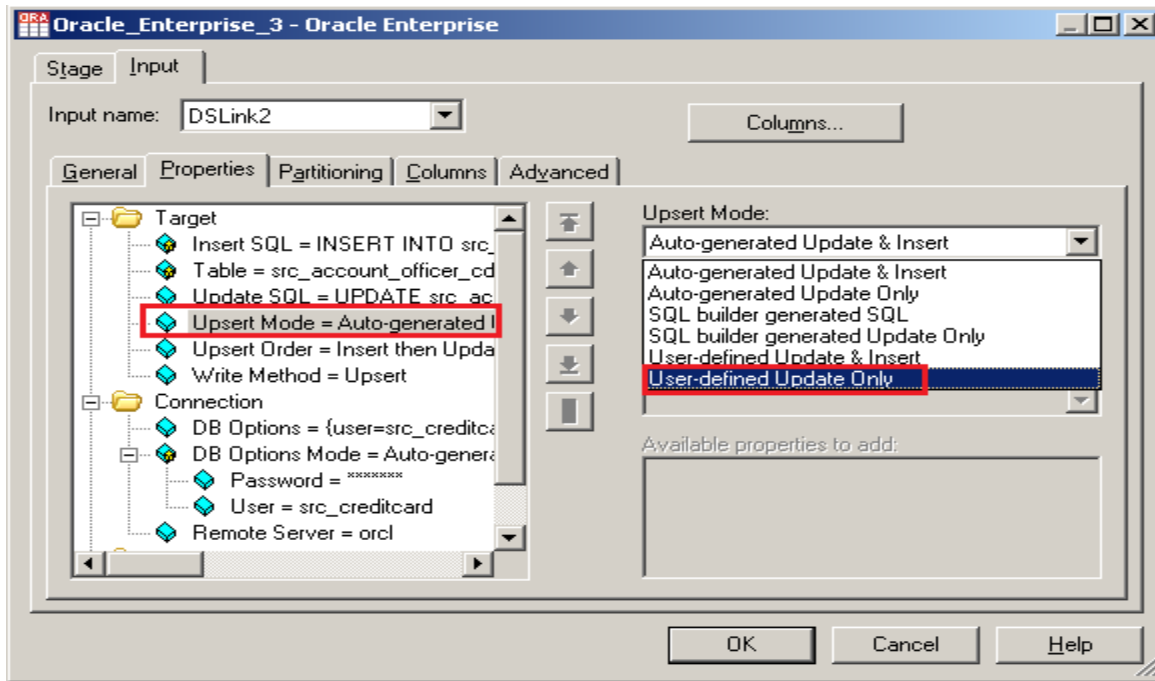
**Step-22:-** Now there are two ways to load data, directly load or user define load. We choose user define that's why we choose upsert.



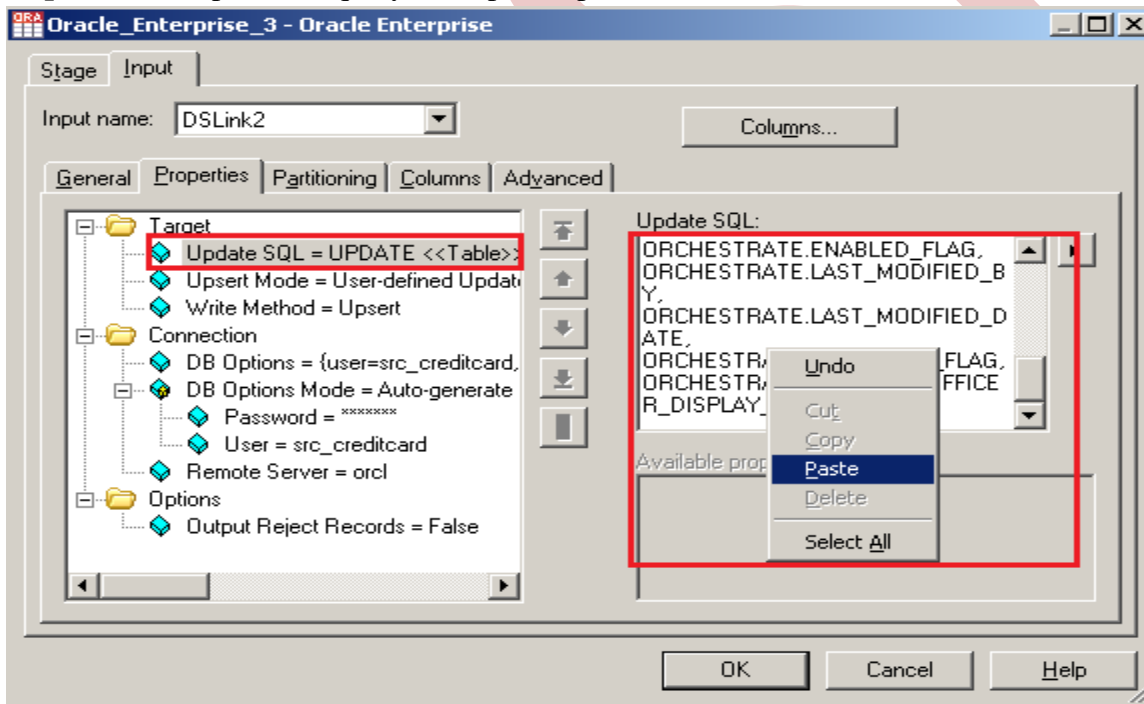
**Step-23:-** See below ,there is insert query and update we want to insert data in oracle so copied the insert query.



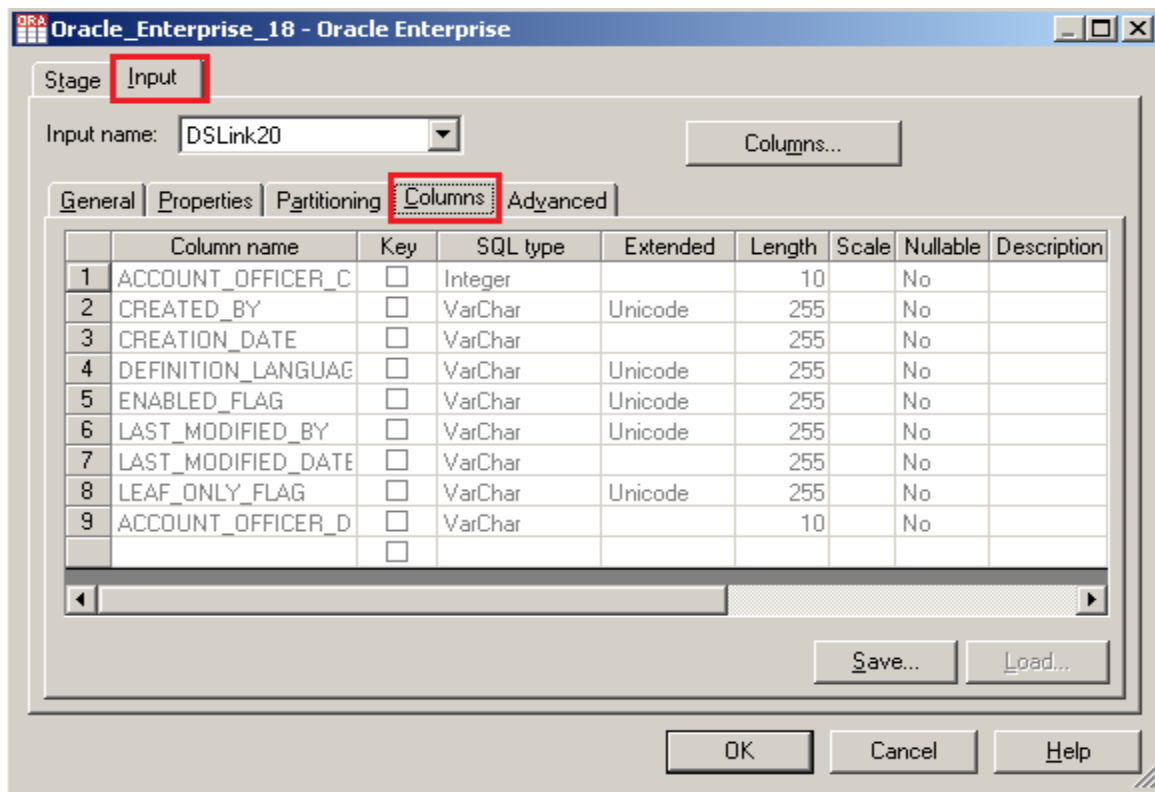
**Step-24:-** We only want to update our data by insert data in table ,so we copied the insert query and put in this and because we only want to update so we have chosen update only method.



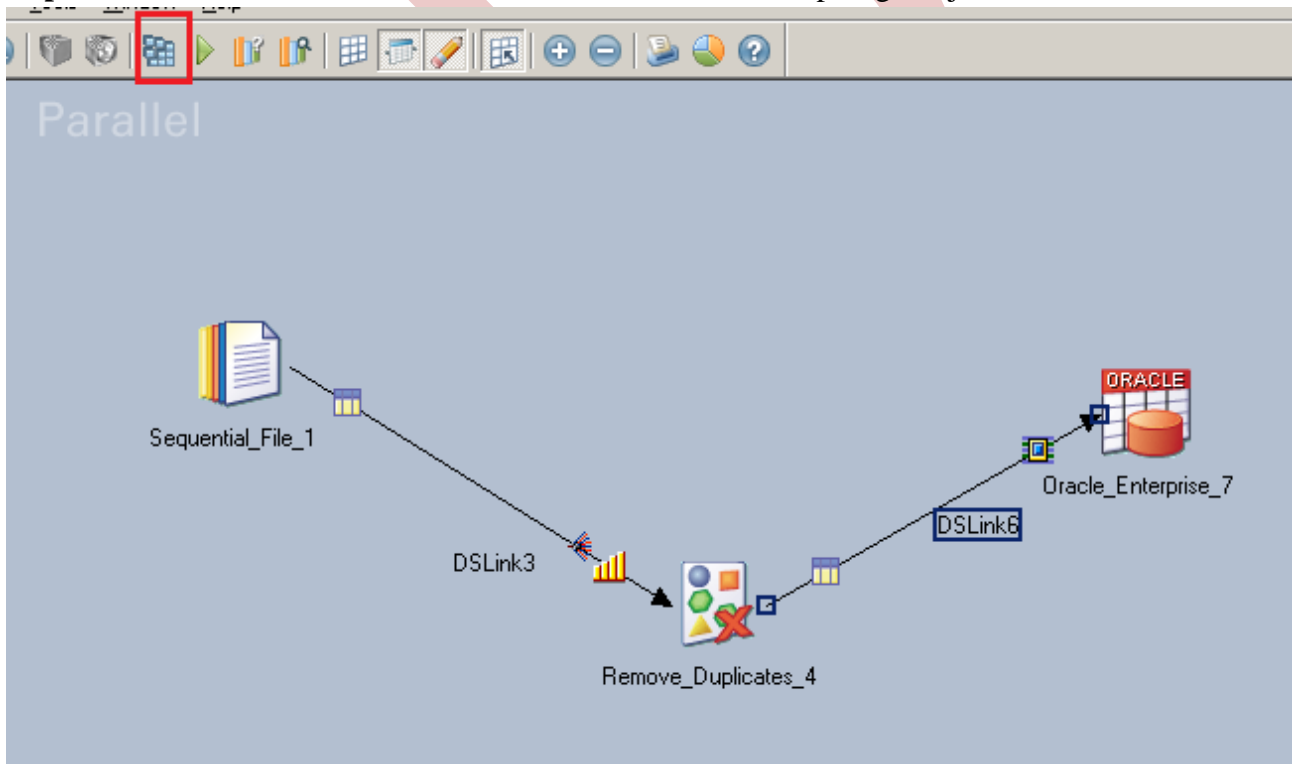
**Step-25:-** Now paste the query into update space.



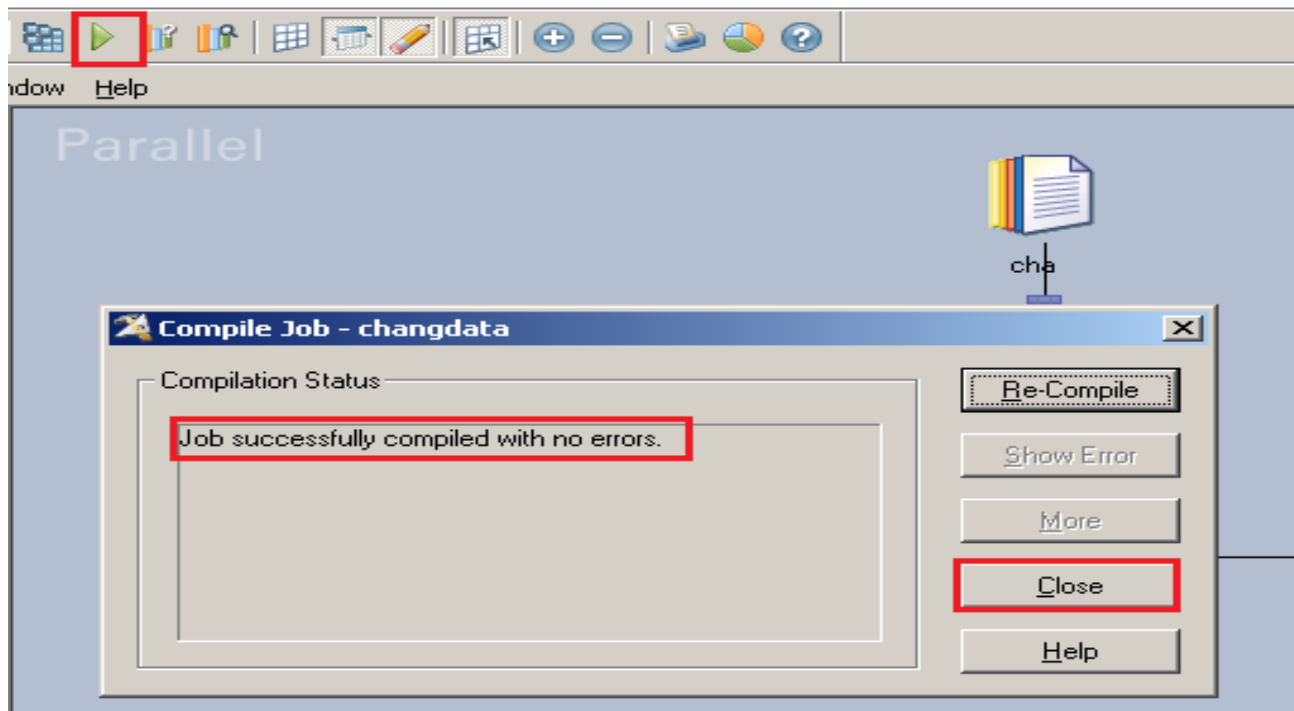
**Step-26:-** Click on column tab and see all the columns are showing or not.



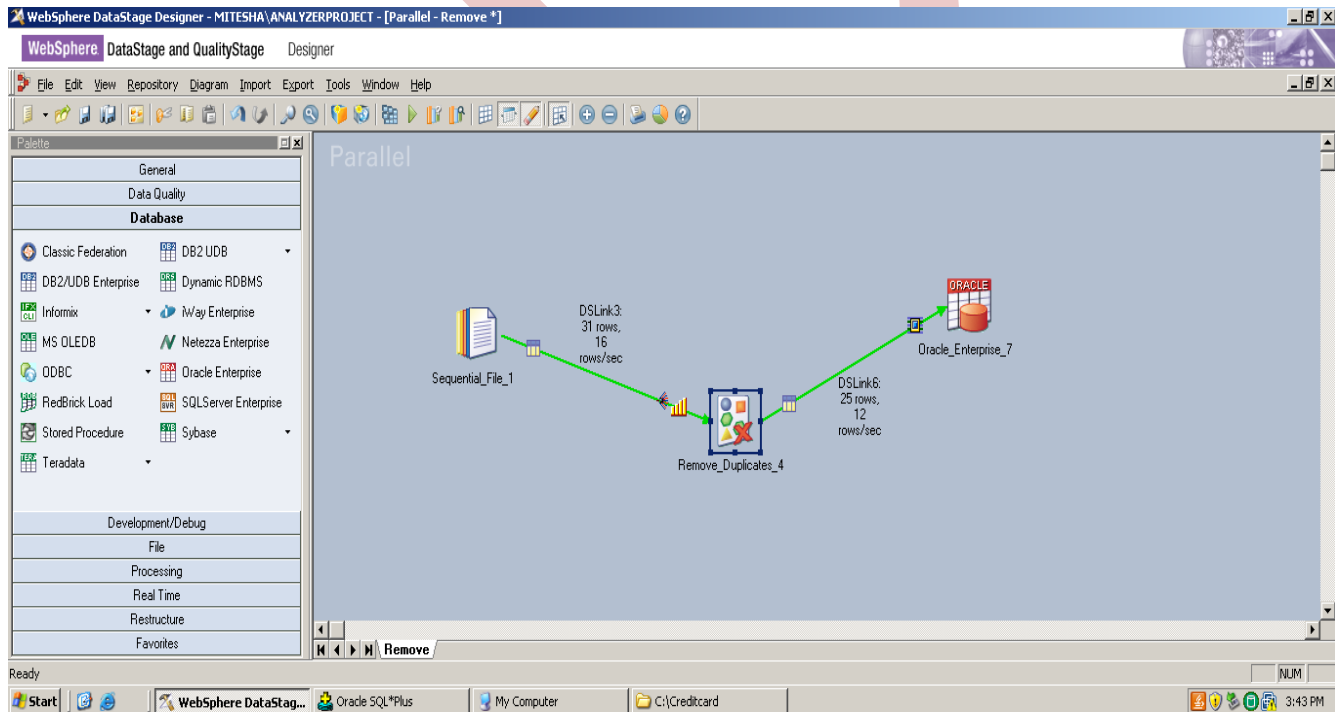
**Step-27:-** This shown like this and now click on button for compiling our job.



**Step-28:-** If this shown that means your compilation is done otherwise it shows error and now click on run button or (CTRL+F5).



**Step-29:-** After click on run wait for a while than it shows GREEN line that means your transformation is successfully done otherwise if it shows RED Line that means not Done and BLUE Line means Under Process.





**Step-30:-** Now Go to SQL Console and connect with the same login credentials as I mentioned in oracle enterprise stage src\_creditcard/password so you can choose your own username and password. Before loading, remember to check structure of this file is available and then you should be able to load the data. then simple query:

```
<SELECT * FROM SRC_ACCOUNT_OFFICER_CD;>
```

	ACCOUNT_OFFI...	CREATED...	CREATION_DATE	DEFINITION_L...	ENABLED_F...	LAST_MODIFIE...	LAST_MODIFIED_DATE	LEAF...	ACCO
1	52223	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	52223
2	52535	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	52534
3	53110	RTUSER	11-JUN-10 07.19.36.0000000...	IND	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	53110
4	53458	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	53457
5	53961	RTUSER	11-JUN-10 07.19.36.0000000...	UK	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	53960
6	54254	RTUSER	11-JUN-10 07.19.36.0000000...	IND	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	54253
7	54608	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	54607
8	54706	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	54705
9	55304	RTUSER	11-JUN-10 07.19.36.0000000...	IND	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	55303
10	55362	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	55362
11	55675	RTUSER	11-JUN-10 07.19.36.0000000...	UK	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	55674
12	56174	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	56173
13	56339	RTUSER	11-JUN-10 07.19.36.0000000...	UK	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	56338
14	56619	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	56618
15	57619	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	57619
16	58010	RTUSER	11-JUN-10 07.19.36.0000000...	IND	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	58010
17	58235	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	58235
18	58616	RTUSER	11-JUN-10 07.19.36.0000000...	US	Y	RTUSER	11-JUN-10 07.19.36.000000...	Y	58616